ISSN: 2321-7782 (Online)
Volume 2, Issue 7, July 2014
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

Automatic Clustering using Feature Partition for Political Analysis

Bhagyashri S. Patil[1]
Rajarshi Shahu College Of Engineering
Pune University, Tathawade
Pune – India

Prof. Rushali A. Deshmukh[2]
Rajarshi Shahu College Of Engineering
Pune University, Tathawade
Pune – India

Abstract: ***Todays World is digital world. The data present on social media like Facebook, Twitters, google plus etc. are in digital format. And the data is in non-standard format. In data mining, text mining fields like computer forensic, in police investigation examining the number of files or messages or comments every day is main task. A difficult task is to group those unstructured data from documents. In document clustering determining the number of cluster is substantial job. Our proposed methodology in order to document features i.e. represent the document by elements that can be considered as discriminative words and discard non-discriminative words. Automatic partition and automatic clustering without requiring the input are the main task. Our approach consists of making the technique semi-supervised or unsupervised such that technique becomes effective and robust than existing approaches. We would also apply this approach to the political news comments data to analyze the political impact of news and incidence. As well as Considering the different combination techniques, combination for additional parameters and analysis with existing techniques such as k-means, k-mediods algorithm i.e. comparative studies.***

Keywords: Document Clustering, pattern recognition, Model-based Clustering, Feature Partition, Comparative Study.

I. INTRODUCTION

Document clustering grouping unlabelled text documents into meaningful clusters is of great interest. Clustering can be metric for unsupervised problem. Document clustering becoming important techniques for unsupervised document organization, database application, automatic topic extraction, filtering the information and fast information retrieval. A web search engine often contains huge datasets. The users have to browse the whole dataset to estimate the number of clusters K. In order to estimate K is time consuming and unrealistic while tackling with huge dataset. The clustering process tends to mislead when improper estimation of K and degrade the performance as the dataset goes on smaller or bigger number of clusters used. The number of clusters K is known before the Document clustering process in traditional Document clustering approach. Whenever we deal with dataset our intention is accuracy of clustering should be maintained. Design the model without consideration of predefined K.

We attempt to design number of clusters K, which is discovered automatically. To partition Documents we are using a Dirichlet process mixture (DPM) model. When the number of clusters is unknown the DPM model shows strong results. Idea of using DPM model is at the same time considers both the data similarity and clustering property of the Dirichlet Process (DP). DPM model is flexible in nature whenever a new data point arrives it either creates from exiting cluster or start new cluster. Each Document contains large amount of words i.e. discriminative words and non-discriminative words. The existence of non-discriminative words confuses the clustering process so only discriminative words are useful. A non-discriminative word leads to poor clustering solution. Our contribution is to resolve the issues of document clustering and use the DPM model to tackle those issues. A DPMFP model is nothing but the extension of DPM model using the feature partition[12]. Each and every Document is mixture of two components they are discriminative and non-discriminative words. The words which are generated

from specific cluster to which document belongs are said to be discriminative words and the words generated from general background shared by all documents said to be non-discriminative words. To infer the latent cluster structure, only discriminative words are useful.

The variational inference algorithm and Gibbs sampling algorithm are used to infer DPM parameters. For document clustering the Gibbs sampling algorithm is difficult to apply because it needs more time to converge. As we know the representation of text documents is high-dimensionally represented so it is even harder to apply. To infer document collection structure in quicker manner we are going to use the variational inference algorithm. However, in our DPMFP approach, we need to infer the document collection structure as well as partition of document words simultaneously. Therefore we cannot directly apply the variational inference algorithm without considering the DPM model. Third approach designs a method to estimate the document collection structure for DPMFP model. To simplify process of parameter estimation a Dirichlet Multinomial Allocation (DMA) model, namely DMAFP is used to approximate the DMAFP. We could conduct experiments on our proposed approach by using real time dataset, real time application. We will also use our approach for analysing political data normally available in daily newspapers in electronics format in the readers comment section of the news. Our approach consists of making the technique semi-supervised or unsupervised such that technique becomes effective and robust than existing approaches.

The overview of our paper as follows: In section II, Literature Survey reviews the related work on document clustering and determining number of clusters. In section III Implementation Details Background knowledge of the DPM model we describe our proposed models DPMFP model and its DMAFP approximation. Our proposed variational inference algorithm and Gibbs sampling algorithm. Section IV shows the result set and data set and section V finally concludes conclusion and future work.

II. LITERATURE REVIEW

The document clustering algorithms and models categorised based predefined input parameters to form the number of cluster. Pros of K-means relatively scalable and efficient in processing large data sets; complexity is O (i k n), where i is the number of iterations, *k* is the number of clusters, n is the number of objects. Normally, k<<n and i<<n. Cons of k-means applicable only when the mean of a cluster is defined; not applicable to categorical data. Need to predefine k, before clustering process. Not suitable to discover clusters with non-convex shape or clusters of very different size. Noisy data and outliers difficult to handle and ends at local optimum. Initial partition leads to how many total runs needed [4]. Pros of K-mediods are more robust than k-means in the presence of noise and outliers; because a mediods is less influenced by outliers or other extreme values than a mean. Cons of k mediods relatively more costly; complexity is *O (i k (n-k) 2)*, where i is the total number of iterations, k is the Number of clusters, and *n* is the number of objects. Relatively not so much efficient. Need to predefine *k* [8]. Fuzzy-means (FCM) and Self-Organizing Maps (SOM) [10] these algorithms have their own properties and widely used in practice. Most existing document clustering methods are based on the Vector Space Model (VSM) [11]. Which are widely used for data representation text classification and clustering? The VSM represents each document as a feature vector of the terms in the document. Each feature vector contains term weights of the terms in the document. A survey of document clustering algorithms with topic discovery is presented in [6]. The existing topic detection method used to compare our proposed work is topic detection by Clustering Keywords [3].

III. IMPLEMENTATION DETAIL AND BACKGROUND

A. Background Design

1. *Dirichlet Process Mixture Model*

The DPM model is kind of infinite mixture model & flexible mixture model. In which the number of mixture component increases as new data arises.

2. Dirichlet Multinomial Allocation model

The DPM model can be derived as the limit of a sequence of finite mixture models when the number of mixture component taken to infinity. Approximation for DPM model is Dirichlet Multinomial Allocation model (DMA) model.

3. Variational and Gibbs sampling algorithm

The variational inference algorithm could be used for document collection clustering in quicker way and Gibbs algorithm is used to infer model based on DP prior, but it's difficult to apply for document clustering its takes long time and even harder to apply when dataset is large. Comparative between variational and Gibbs sampling algorithm variational inference algorithm is used due to faster execution less time it requires comparative to Gibbs algorithm.

B. Mathematical model and Algorithm

Input set:

w= {1, 2,….,W} word item from a vocabulary. (Find the different words from the entire document (training))

A document x is represented as a W dimensional vector xd= {xd1, xd2, ..., xdw} (Feature Vector Dimension for each document is W)

Where, xd is the number of appearance of word w of the document xd. X={x1, x2….xD} (count the frequency of every word from set, w, in each document)

Where X is collection of D document.

V= {v1, v2… vw} (Threshold the each element of X with typical values to get binary vector)

A where V is latent binary vector to partition document words into two group's discriminative and non-discriminative words.

Let denote Ω the discriminative word set. Words not belong to Ω are consider as non-discriminative words. (We get output as follows)

V= 1, if w exist in Ω than discriminative word , Otherwise its 0, if w not exist Ω than non-discriminative word.

IV. RESULTS AND EXPERIMENTAL

A. Dataset: News dataset

We pre-processed the data sets by stop word removal. High-frequency and Low-frequency words were removed. The purpose of such processing is to eliminate those words which obviously unable to define the latent cluster structure. Thresholds for removing high-frequency and low-frequency words for data sets were set 100 and 1, respectively. To evaluate the performance of this proposed approach, we use a social media review dataset, which includes huge review articles. The proposed approach for evaluating the discriminative words and clustering is an automatic method that is independent of predefined number of clusters, we use the entire dataset. Result obtained for General news dataset as Total Words: 6572,Unique Words: 1352,Filtered words: 707, Frequency range low to high 1 to 100.

B. Result set: Real-life political comments

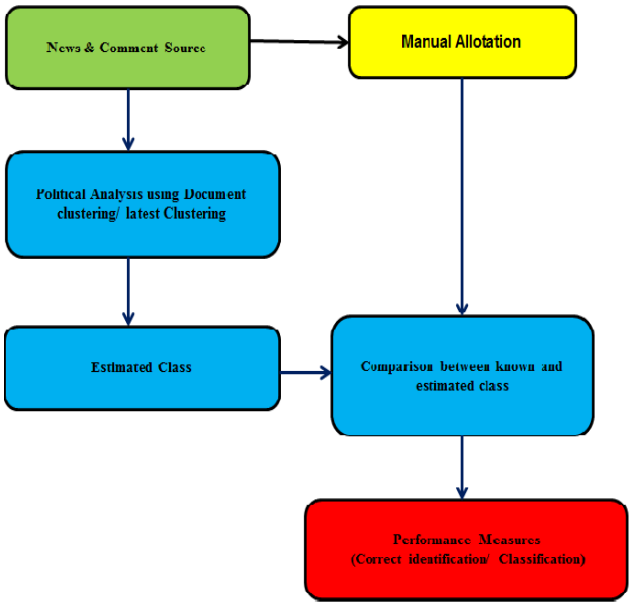

Fig1: Proposed data flow diagram for clustering of political data

Fig 1 depicts more and more internet user prefer to comment on social hot issues like politics today and their views become very useful for government decision-making. The news and related comments often confuses to take decision. However, it becomes a key problem to analyse them automatically in order to provide references for decision-making. One of effective way is to cluster news comments. In this paper, we discuss the clustering algorithm and how to cluster news comments in order to obtain types of a special news comments. We do an experiment on a real dataset collected from the news recommender system we developed for government decision-making. Primary results are shown that our clustering method is effective and can be taken as an analysis method used in our recommender system. Both the variational inference algorithm and the blocked Gibbs sampling algorithm were used to infer the cluster structure as well as the partition of features for the data set.

To do document clustering using VDPMM.

V. DESCRIPTION

The document clustering using various functionalities of VDPMM. Steps are as follows:

1. Extract comments from web-pages and store them into “temp” folder.
2. To remove some html tags from extracted comments in temp folder we have created another folder named “dataset” and stored HTML tags removed comments in that folder.
3. Extract words from the each comment file which are stored in dataset folder and write “words.txt” file containing words from all the comments.
4. Filter the words saved in words.txt file based on their frequency and store them in “filtered_words.txt” file.
5. Calculated feature vector for each file in dataset folder by matching words from comments file with filtered words.
6. To do clustering implement functionalities of VDPMM in which clustering is done automatically and we don’t have to give number of clusters as input.
7. To cluster all files in dataset folder, apply functionalities of VDPMM on the feature vector of all files by passing feature vector as an argument to vdpgm function.
8. This function calculates number of clusters for the given feature vector.
9. Find the files which are allocated to particular cluster using ED of feature vector and centroid of cluster calculated using VDPMM.

Results for Kmeans & after implementaion of our propsed model:

- Fig 2 depicts that, K-means after entering number of clusters 4:

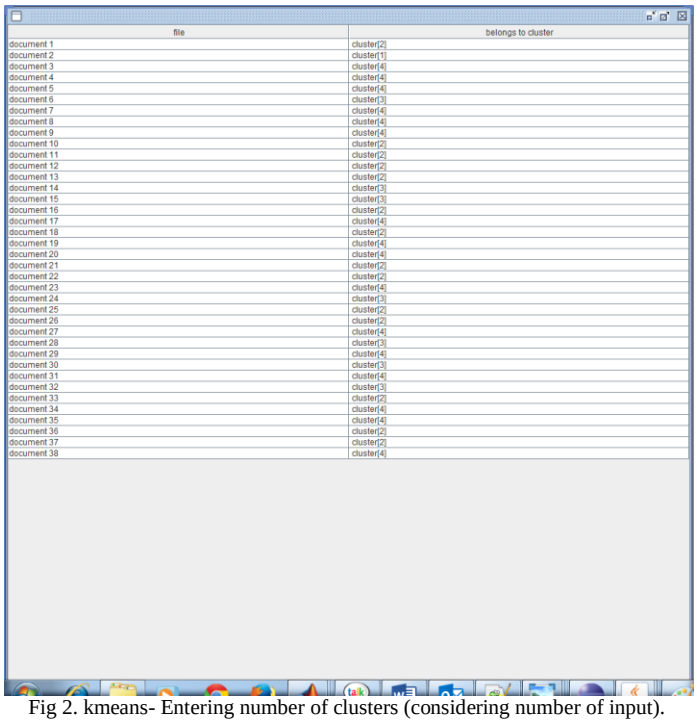

Fig 2. kmeans- Entering number of clusters (considering number of input).

- Fig 3 depicts that, after implementaion of our approach that is without considering number of inputs.

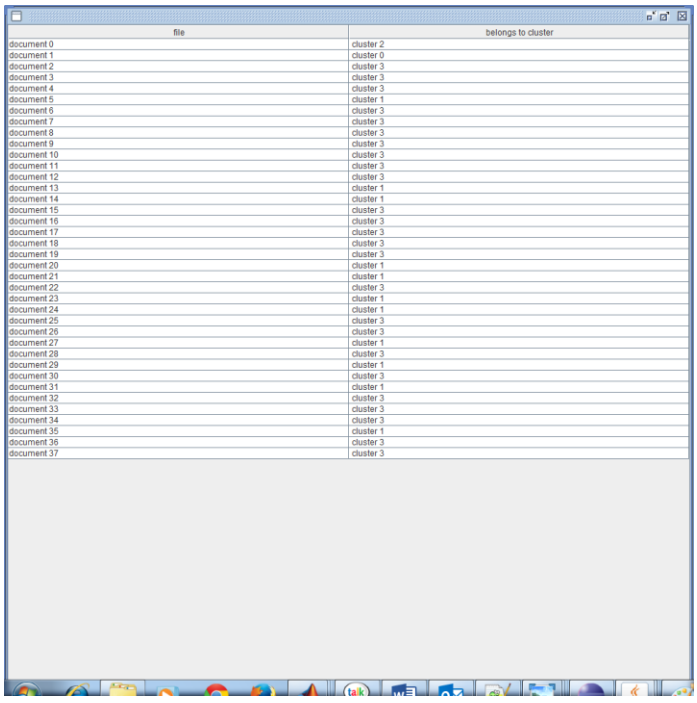

Fig 3. VDPMM (without considering number of inputs)

VI. CONCLUSION

Our proposed DPMFP approach handles document clustering and feature selection simultaneously. We constrain the DPM model only to define the cluster structure of the data with discriminative features which are identified by a latent binary vector. Our experiment shows that DPMFP approach groups document dataset into meaningful clusters without requiring the number of clusters known in advance. The analysis and comparison between our approach and other approaches could conclude that our approach is more effective and robust .We also applied approach to the political comments data. To improve the performance of

our approach the additional information could be used to select good model parameters and our model select more precise discriminative words set.

References

1. K. Jain and R. C. Dubes, Algorithms for Clustering Data. Englewood Cliffs, NJ: Prentice-Hall, 1988.
2. Aggarwal, C. C. Charu, and C. X. Zhai, Eds., "Chapter 4: A Survey of Text Clustering Algorithms," in Mining Text Data. NewYork: Springer, 2012.
3. Christian Wartena and Rogier Brussee, "Topic Detection by Clustering Keywords", IEEE 19th International Conference and Expert System Application, 2008.
4. D T Pham_, S S Dimov, and C D Nguyen "Selection of K in K-means clustering" Proc. IMechE Vol. 219 Part C: J. Mechanical Engineering Science.
5. Information Theory, Inference, and Learning Algorithms David J.C. MacKay Copyright Cambridge University Press 2003.
6. Jayabharathy, S. Kanmani and A. Ayeshaa Parveen, "A Survey of Document Clustering Algorithms with Topic Discovery", Journal of Computing, Vol. 3, Issue 2, February 2011.
7. Jiawei Han and Micheline Kamber, "Data Mining Concepts and techniques ", Second Edition.
8. Kurt Hornik , Ingo Feinerer, Martin Kober, Christian Buchta "Spherical k-Means Clustering" Journal of Statistical Software September 2012, Volume 50, Issue 10.
9. Pattern Recognition and Machine Learning Christopher M. Bishop Copyright c 2002–2006.
10. S. Haykin, Neural Networks: A Comprehensive Foundation. Englewood Cliffs, NJ: Prentice-Hall, 1998.
11. Salton and M. J. McGill, Introduction to Modern Information Retrieval. McGraw-Hill, 1983.
12. Z. Wang, and G. Yu, R. Huang, "Document Clustering via Dirichlet Process Mixture Model with Feature Selection," Proc. ACM Int'l Conf. Knowledge Discovery and Data Mining, pp. 763-772, 2010.

